DOCUMENT RESUME

ED 268 133 TM 850 732

AUTHOR Wainer, Howard
TITLE On the Study of Matching Cut-Scores to Test Characteristics: An Observed Score Approach. Program Statistics Research Technical Report Series.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO ETS-RR-85-14; ETS-TR-85-56
PUB DATE May 85
NOTE 9p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; *Item Analysis; *Latent Trait Theory; Mastery Tests; Raw Scores; Scores; *Scoring; *Test Reliability; Test Validity
IDENTIFIERS *Cut Score Characteristic Function

ABSTRACT

Techniques derived from item response theory are useful for estimating the reliability of test classification above and below the cutting score. Test developers can construct a test whose information is peaked in the region of the cutting score; users can select a test which provides the most information in this region. The Cut-Score Characteristic Function (CSCF) mirrors the item response theory approach and can be used with observed scores. The test is arbitrarily divided in half, and the cut score is relocated. With a 100-item test having a cut score of 60, two 50-item tests are created, each having a cut score of 30. The scores for the first test are plotted on the x-axis of a graph, while the proportion who passed the second test is plotted on the y-axis. The resulting curves indicate a test (1) with high discriminating ability whose location matches the cut score reasonably well; (2) which matches the cut score, but does not discriminate as well; or (3) which discriminates well but is poorly matched to the cut score. This method emphasizes the test's quality in the region of the cut score. A test could be split into halves repeatedly. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

On the Study of Matching Cut-Scores to Test Characteristics: An Observed Score Approach

Howard Wainer

Program Statistics Research
Technical Report No. 85-56

Research Report No. 85-14

May 1985

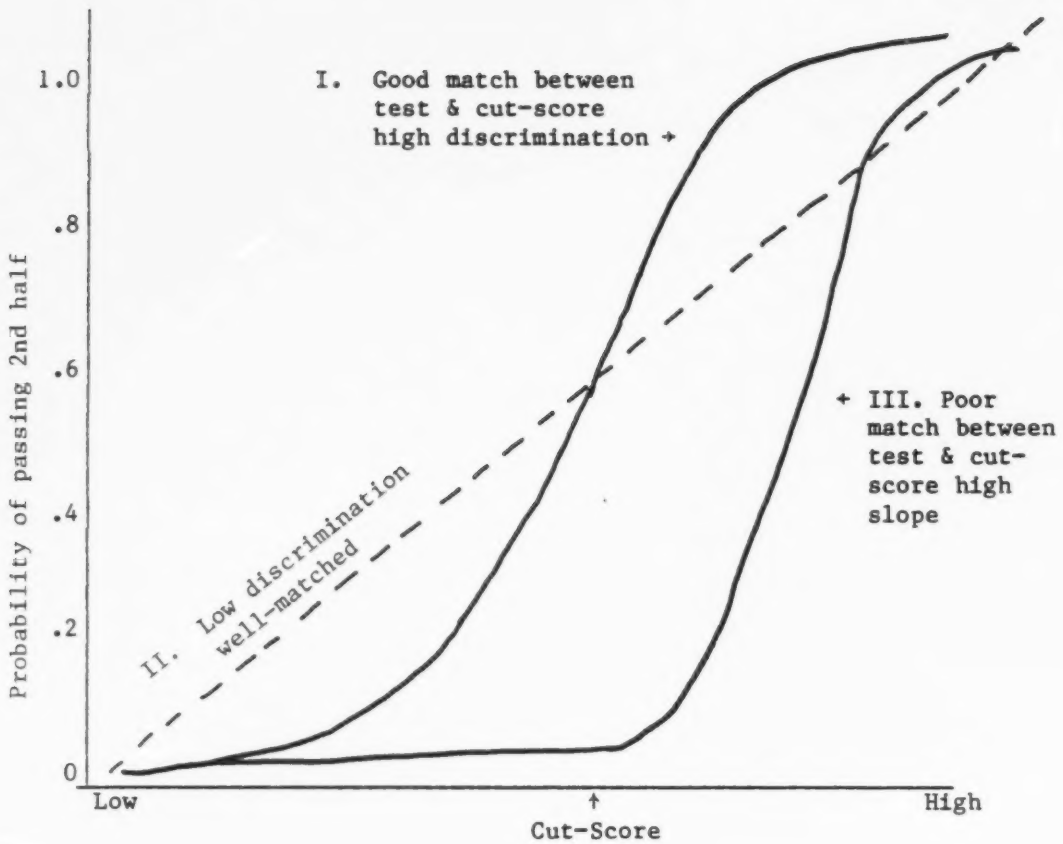
Copyright © 1985 by Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

When a test is used with a cut score, the usual measures of test reliability (i.e. KR20) that have as their basis the internal consistency of the test, are often inappropriate. This is because they characterize the structure of the entire range of the test, rather than in the immediate vicinity of the cut-score. What is needed is some measure of the reliability of classification above and below the cut-score. It seems to me that the most graceful way of estimating this is through the use of test information curves derived from Item Response Theory (IRT). Such curves provide a measure of the information available for judgments at each point on the ability continuum. Test developers can utilize such curves by constructing a test whose information is peaked in the region of the cut-score; whereas prospective test users can examine the information curves of a variety of candidate tests to choose the one that has the most information in the region of the cut-score.

In many situations IRT has not yet been implemented either because of technical or political reasons. It seems sensible to try to develop a measure that mirrors the IRT approach that can be used with observed scores. One notion that seems reasonable is to construct what we will call the Cut-Score Characteristic Function (CSCF). To do this, suppose we arbitrarily divide the test in half (say odd items vs. even items), and similarly relocate the cut score. Thus, if we have a 100 item test with a cut score at 60, we now consider a two tests of 50 items each with a cut-score at 30. We then make the following plot:



We plot the scores for the (arbitrarily chosen) first half of the test on the x-axis and the proportion who passed the cut-score on the second half of the test on the y-axis. The resulting curves can have a variety of shapes; some of which are illustrated above.

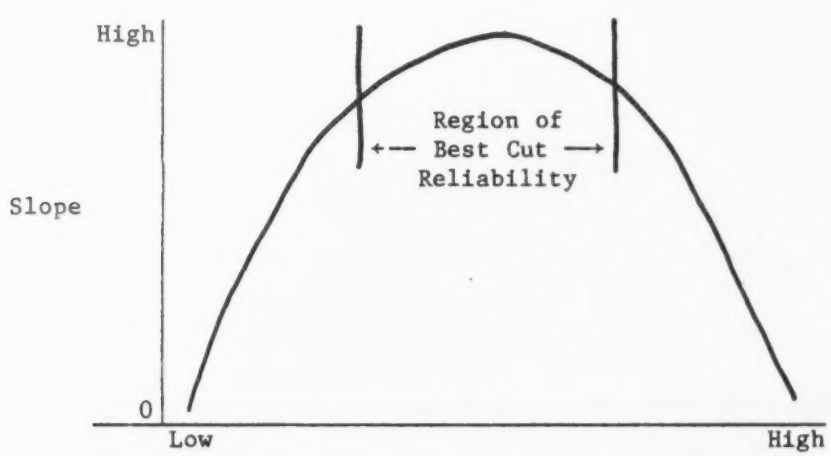
I. A test with a rather high discriminating ability (a steep slope) whose location matches the cut-score reasonably well.

II. A test which matches the cut-score chosen, but which does not discriminate as well.

III. A test which discriminates well but which is poorly matched to the cut-score (such an outcome is probably caused by a poor splitting technique originally).

The principal comparison should be between test curves of type I and II. We have opted for displaying the entire curve rather than dividing it up into the traditional 2x2 table (pass-fail on 1st half vs. pass-fail on 2nd) so as to avoid over-influencing our decisions by the 'easy' cases (i.e. those individuals who are far from the cut-score on both halves). It seems to me that the real mark of the quality of a test is how well it performs in the region of the cut-score. Thus we have preserved this distinction.

We believe that through the examination of such curves one can readily compare the efficacy of two tests at a given test score. In addition, since there is a (conceivably) different curve for each test at each cut-score one could easily produce a variety of curves for the same test to see at what cut-score that test performs best (and worst). This can be automated by fitting a smooth curve to this function (say a logistic function) and then plotting the slope of this curve at each cut-score. A result might look like this:



One could then develop an index that would characterize the quality of any particular test cut-score combination. To do this we establish an ideal test in which the CSCF is zero below the cut-score and 1 above it, and measure the deviation of the actual CSCF from it. More precisely, let us define:

x = observed score on half the test

c = cut-score

$$I(x) = \begin{cases} 0 & x < c \\ 1 & x > c \end{cases}$$

$F(x)$ = the CSCF for that test and that cut-score

$w(x)$ = proportion of individuals at score x

A measure of the quality of any test/cut-score combination then would be:

$$E_1 = \sum_x [F(x) - I(x)]^2 \, w(x)$$

or, equivalently

$$E = \sum_{x<c} [F(x)]^2 \, w(x) + \sum_{x>c} [F(x) - 1]^2 \, w(x)$$

Of course this is a one-sided view, and we could reverse what we call the 1st half and 2nd half and do the same thing. This would provide us with two estimates of the same statistic, say E_1 and E_2, which we might wish to concatenate into a single number that would represent that test, cut-score and test division. An easy combination would be $E = (E_1 + E_2)/2$.

Obviously, for every sampling plan that divides the test in half we would have another value of E, and we could randomly split the test into two halves many times and calculate the value of E each time. This could provide

us with a direct measure of the variablity of E. While we have not tested such a notion, it is attractive to think the stability of this statistic can can be measured with such a bootstrapping method. The bounds calculated from this bootstrap would then provide us with a measure of how much the quality of the observed measure was due to the vagaries of item sampling.